

# Visualizing The Central Limit Theorem

BY MADHURI S. MULEKAR

## Abstract

For students in an introductory statistics course, the probabilistic ideas involving sampling variation are difficult to understand. This paper describes the use of technology for teaching the ideas behind the Central Limit Theorem (CLT) to students in a non-calculus based, introductory statistics course.

## What Do We Want Students to Know About the CLT?

Increasingly, computers and graphing calculators are being used in introductory statistics courses. However, such use of technology has been mostly limited to mastering the data analysis technique. The problem faced by teachers of introductory courses is that Statistics is not an intuitive subject. Many statistical concepts are probabilistic and many students have math barriers.

One possible solution to this problem is visualization of concepts. Concepts presented visually are more concrete. Advances in technology have made it possible to present statistical ideas visually, and make concepts concrete by allowing students to conduct statistical experiments.

The CLT is one of the concepts taught in introductory statistics courses. Typically, the introduction of the concept of sampling distribution leads, naturally, to the concept of CLT. Consider the following statement of the CLT given by Serfling (2001), a leading

book on the approximation theorems:

Let  $X_i$  be I.I.D. with mean  $\mu$  and finite variance  $\sigma^2$ .

Then, as  $n \rightarrow \infty$ ,

$$\sqrt{n} \left( \frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \rightarrow N(0, \sigma^2).$$

$$\text{That is, } \frac{1}{n} \sum_{i=1}^n X_i \text{ is } AN \left( \mu, \frac{\sigma^2}{n} \right).$$

Clearly, this definition is couched in the concepts and symbols of calculus. No doubt, students will find such a definition intimidating. Since Serfling (2001) is not among the textbooks used for introductory, non-calculus-based courses, let us consider one of the more popular textbooks from that category. The statement of the CLT found in Yates, et al (2003), is wordier but involves fewer notations. It is given below:

Draw a SRS of size  $n$  from any population whatsoever with mean  $\mu$  and finite standard deviation  $\sigma$ .

When  $n$  is large, the sampling distribution of the sample mean  $\bar{x}$  is close to the normal distribution,

$N \left( \mu, \frac{\sigma}{\sqrt{n}} \right)$ , with mean  $\mu$  and standard deviation  $\frac{\sigma}{\sqrt{n}}$ .

Here, "SRS" refers to the simple random sample. In spite of the simplified statement format, the result of this theorem makes an intuitive appeal to the notion of limit from calculus, and teachers are faced with the difficult task of explaining it to students who have no calculus background. As a result, students have difficulty understanding the meaning and usefulness of this important theorem. What is it that we expect students to know about this theorem? Do we want them to memorize the statement of the theorem and reproduce it? No. We want them to understand the underlying ideas and apply them to problems in real life situations. The CLT is a complex mixture of several concepts. In traditional textbooks it is covered somewhere around chapter 6 or 7. Several concepts introduced in earlier chapters, such as the following, come together in the CLT.

- (1) Population versus sample,
- (2) Population mean versus sample mean,
- (3) Random sampling,
- (4) Variability or variation,
- (5) A pattern (distribution) of all possible values that we could possibly observe for the sample mean,
- (6) A population distribution versus sampling distribution,
- (7) Effect of sample size on the variability of sample means,

- (8) Diminishing difference between the actual sampling distribution and the limiting normal distribution,
- (9) Distribution in limit.

Do we realistically expect students without any background in calculus to understand the concepts of convergence of distribution? Certainly not. This being the case, we must decide what we DO expect them to understand after studying the CLT. As described by Newton and Harvill (1997), the basic objectives behind teaching the CLT are to make students aware of the following:

- The shape of the distribution of many sample means, each calculated from a different, randomly selected sample, is an approximately normal (or bell-shaped) curve.
- The mean of the sampling distribution of the sample mean is the mean of the parent distribution (the mean of the means is the mean).
- As the sample size gets larger, the sampling distribution of the sample mean becomes more symmetric.
- As the sample size gets larger, the sampling distribution of the sample means becomes narrower.
- For more skewed parent distributions, larger sample sizes are needed in order for the sampling distribution to become approximately normal.
- There is no such magic number as 25, 30, 50, or 100, as suggested by so many textbooks, above which the approximation is adequate.

### How Do We Demonstrate the CLT?

Simulation along with dynamic graphics can be used to produce moving visualizations of concepts that are difficult to understand without proper mathematical background. Highly interactive graphic displays of simulation results, used in classroom instruction, as well as carefully designed lab experiments, can help students understand the concepts.

There are several software packages available on the market (some of which are share-ware) from which instructors can choose suitable software, based on such considerations as the availability of equipment and transportability. For example:

- StataQuest (Duxbury)
- ActiveStats (Addison-Wesley)
- Visual Statistics (McGraw-Hill Irwin)
- CyberStates (Duxbury)

There are also several web-based applets that can be used to demonstrate the CLT, many of which can be found at the following web addresses:

- [www.stat.vt.edu/~sunder/java/applets/CLT2Applet.htm](http://www.stat.vt.edu/~sunder/java/applets/CLT2Applet.htm)
- [www.ruf.rice.edu/~lane/stat\\_sim/index.html](http://www.ruf.rice.edu/~lane/stat_sim/index.html)
- [www.stat.sc.edu/~west/javahtml/CLT.html](http://www.stat.sc.edu/~west/javahtml/CLT.html)
- [www.math.csusb.edu/faculty/stanton/probstat/CLT.html](http://www.math.csusb.edu/faculty/stanton/probstat/CLT.html)
- [web.uvic.ca/~esfchan/stats/](http://web.uvic.ca/~esfchan/stats/)
- [www.uses.ece.gatech.edu/users/gtz/java/CLT/](http://www.uses.ece.gatech.edu/users/gtz/java/CLT/)

Often, students fail to understand the process behind these applets and feel like numbers are being pulled out of a black box. It is important that they understand the sampling process. However, actually sampling from the population by hand is very time consuming and quickly becomes a repetitive and boring process. Therefore, simulation of the sampling process using a computer program or a scientific calculator is a fitting solution. This paper will demonstrate the use of *Minitab* to illustrate the essential ideas of the Central Limit Theorem, as described above.

#### Using Minitab to Illustrate the CLT:

- **Cal** → **Random data** gives a list of several distributions. Select a distribution. Start with a symmetric distribution like a normal distribution (with any mean and standard deviation) or a binomial distribution with  $p = 0.5$ . Select 100 samples (i.e. 100 rows of data), each of size  $n = 5$  (i.e. save in 5 selected columns, say, C1-C5). Fig. 1 shows how to generate random samples from a symmetric binomial distribution using Minitab. This will result in a table of data consisting of 100 rows and 5 columns. Each row constitutes a sample.
- Compute the mean of each sample using **Cal** → **Row statistics** and save it in another column (say, C6). Fig. 2 shows how to calculate sample means using the row statistics option in Minitab. This will result in 100 sample means.
- Make a dotplot or a histogram to get the sampling distribution of  $\bar{X}$ .
- Compute numerical summary statistics.
- Describe the sampling distribution in terms of the center, spread, and shape of the distribution.
- Repeat the process using  $n = 10, 20, 25, 30$ .
- Describe the effect of sample size on the sampling distributions in terms of the center, spread, and shape of the distribution.

- Then repeat the process with more skewed distributions like a binomial distribution with  $p$  close to 0 or 1, an exponential distribution, or a chi-square distribution.

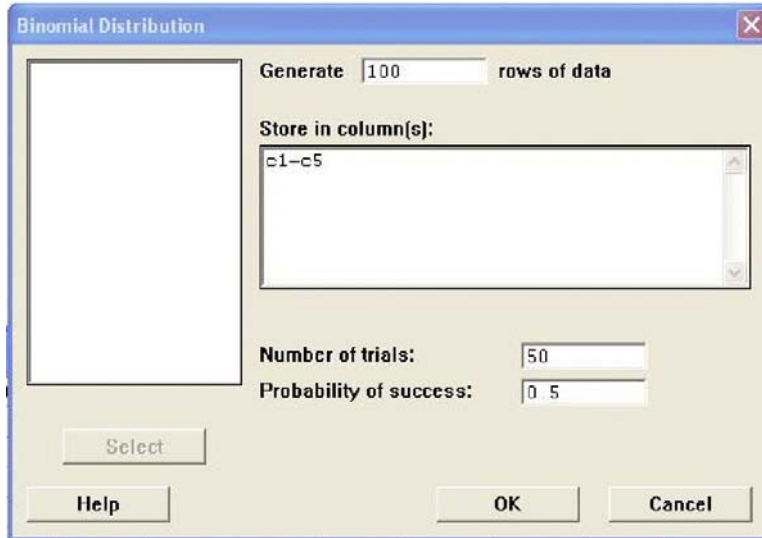


Figure 1: Generating random samples from a binomial distribution

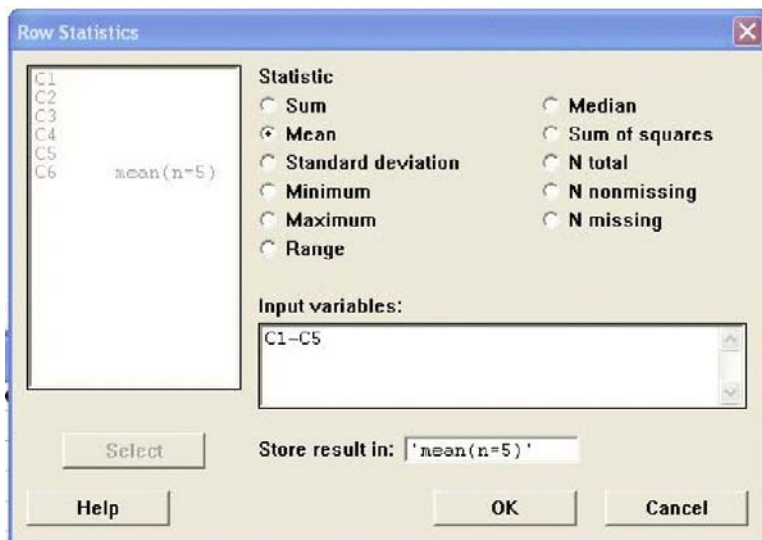


Figure 2: Calculating sample means and saving in a different column

**Example 1.** Here we describe the results of one such simulation. Suppose we selected 200 samples, each of size  $n = 30$ , from an exponential distribution with mean 10. The probabilistic model

for this population is given by:

$$f(x) = \begin{cases} \frac{1}{10}e^{-\frac{x}{10}} & x > 0 \\ 0 & \text{elsewhere.} \end{cases}$$

It is described in Fig. 3. Students in an introductory course are not likely to draw much insight from the formula, but can certainly see from the graph (Fig. 3) that the exponential distribution is far from symmetric.

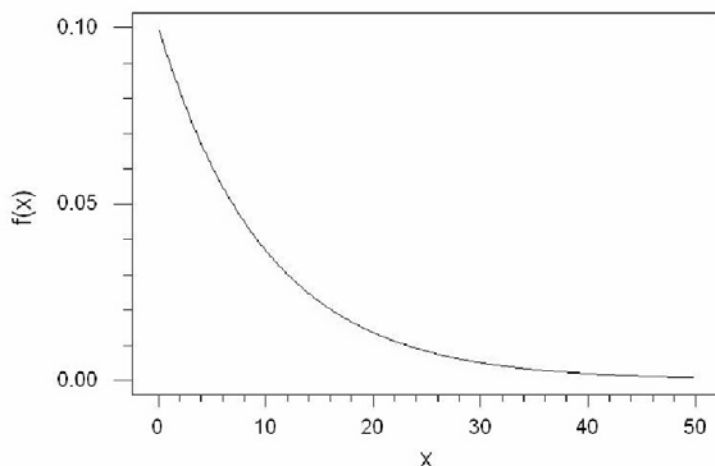


Figure 3: Exponential distribution with mean 10

For this model, the mean is  $\mu = 10$  and the variance is  $\sigma^2 = (10)^2$  (i.e. the standard deviation is  $\sigma = 10$ ). The sample mean  $\bar{x}$  was calculated for each of the 200 samples.

- The average of the 200 sample means turned out to be 9.932, which is reasonably close to the population mean 10.
- The standard deviation for the 200 values of  $\bar{x}$  was calculated and found to be 1.832. Theoretically, the standard deviation of  $\bar{X}$  is  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{30}} = \frac{10}{\sqrt{30}} = 1.825$ , which is not far from the observed value of 1.832.
- A relative frequency histogram for the 200 values of  $\bar{x}$  is shown in Figure 4. It looks mound-shaped and fairly symmetric (with a couple of outliers on the higher end), although we sampled from a very skewed distribution. The normal distribution superimposed on the histogram (Fig. 4) shows that the sampling distribution is closely approximated by the normal distribution.

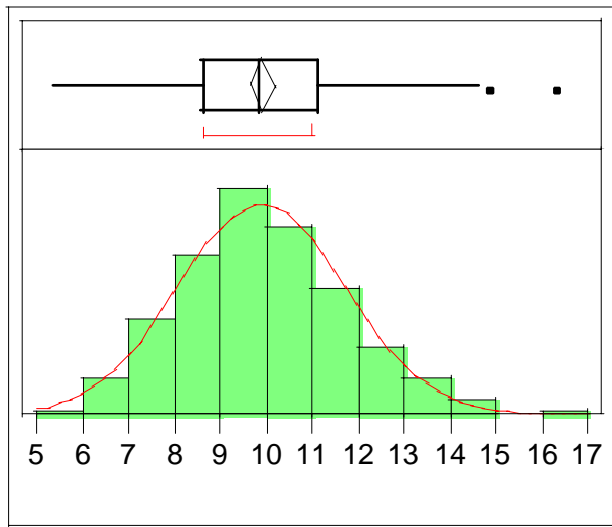


Figure 4: Frequency histogram for sample mean ( $n = 30$ )

To see what happens to the sampling distribution of  $\bar{x}$  for smaller sample sizes, let us take 200 samples of sizes 5 and 10. The sampling distributions of  $\bar{x}$  from one such simulation of 200 samples of sizes 5, 10, and 30 are shown in Fig. 5 below.

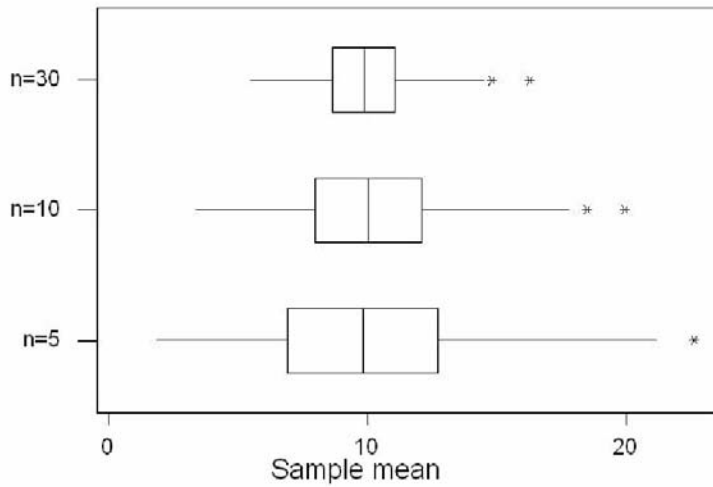


Figure 5a: boxplots

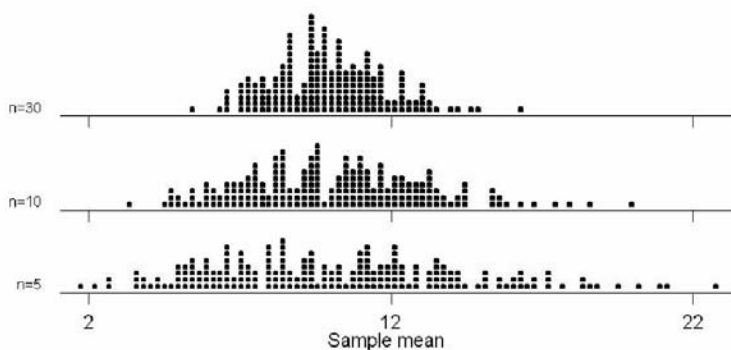


Figure 5b: dotplots

Figure 5: Sampling distributios of  $\bar{x}$  for  $n = 5, 10,$  and  $30$ 

Comparison of three sampling distributions shows that the behavior of the sampling distribution differs for different sample sizes. However, we would like our students to notice a trend in this behavior. Although the population distribution is skewed,

- The mean of the sampling distribution remains close to 10 (the population mean), regardless of the sample size.
- The spread (width) of the sampling distribution is reduced as the sample size is increased.
- The shape of the sampling distribution becomes more symmetric and mound-shaped as the sample size is increased.

The summary statistic from these simulations is reported in Table 1. It supports the conclusions drawn from the graphs of the sampling distributions.

- The average of 200 sample means stays right around 10, the population mean. It does not change with the sample size.
- The standard deviation decreases as the sample size is increased. It can be easily seen that the standard deviation is very close to  $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{n}}$ .
- The shape of the distribution is not easy to determine from the numerical summary. However, a comparison of mean & median, as well as mean and  $Q_1$  and  $Q_3$ , indicate symmetry of the distribution for larger sample sizes.



Variable	N	Mean	Median	Tr Mean	St Dev	SE Mean
mean (n = 5)	200	10.152	9.870	10.014	4.147	0.293
mean (n = 10)	200	10.151	10.053	10.016	2.979	0.211
mean (n = 30)	200	9.932	9.876	9.890	1.832	0.130

Variable	N	Min.	Max.	Q <sub>1</sub>	Q <sub>3</sub>
mean (n = 5)	200	1.830	22.685	6.909	12.760
mean (n = 10)	200	3.353	19.982	8.000	12.130
mean (n = 30)	200	5.444	16.307	8.666	11.098

Table 1: Summary statistic of sample means from an exponential population

We can continue this process with different distributions to show students the meaning of the CLT. Also, we can continue sampling with  $n = 30, 40, 50$ , and show how the sampling distribution gets closer to the normal distribution. Although many perceive a sample size of 30 as a magic number for applying the CLT, we can use these simulations to show that there is no magic number for which the CLT holds under all circumstances. For skewed distributions, multimodal distributions, and distributions with outliers, sample sizes larger than  $n = 30$  are probably needed to see the results of the CLT.

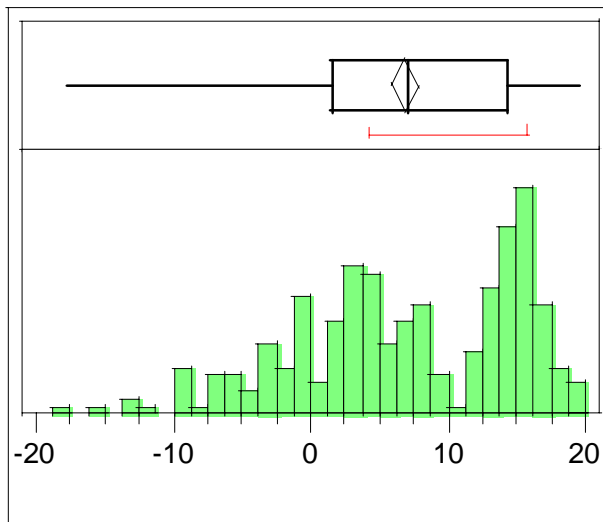


Figure 6: Asymmetric population distribution

**Example 2.** Consider the population described by the distribution in Fig. 6. Clearly, the distribution is left-skewed and possibly

bimodal. The mean and the standard deviation, respectively, for this population are  $\mu = 6.96$  and  $\sigma = 8.08$ . Let us take samples from this population and see how the sample means behave.

Fig. 7a and 7b, respectively, show the sampling distributions of means for samples of sizes  $n = 30$  and  $n = 40$ . from this population. It is obvious that even at sample size  $n = 30$ , the sampling distribution has not attained the symmetric unimodal nature of the normal distribution. The sampling distribution is still somewhat skewed. The sampling distribution for  $n = 40$  is more compact, or less spread, as expected. It is still a little bit left-skewed, but slightly more symmetric than that for  $n = 30$ .

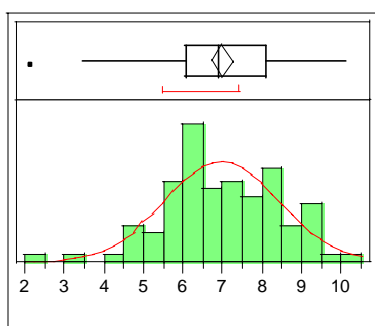


Figure 7a:  $n = 30$

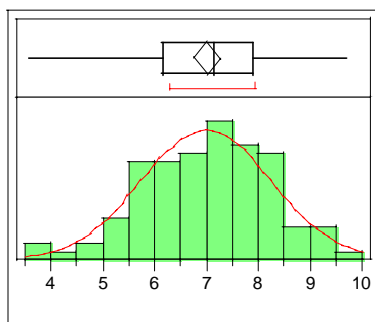


Figure 7b:  $n = 40$

Figure 7: Sampling distribution of the mean for samples from the distribution in Fig. 6

On the other hand, if we are sampling from a fairly symmetric and mound-shaped distribution, the result of the CLT holds even for small samples.

**Example 3.** Let us consider a symmetric population of a binomial experiment with 50 trials and the probability of success 0.5 ( $n = 50$  and  $p = 0.5$ ), as displayed in Fig. 8. Take 200 random

samples of size 5 from this population. The distribution of means of samples of size 5 from this symmetric binomial is displayed in Fig. 9. It shows that the distribution of sample means is symmetric and mound-shaped (very close to the normal distribution), even for samples of size 5.

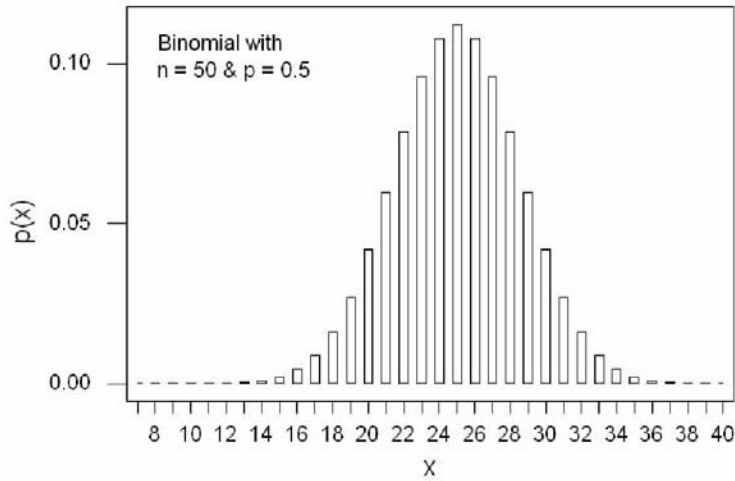


Figure 8: Binomial distribution with  $n = 50$  and  $p = 0.5$

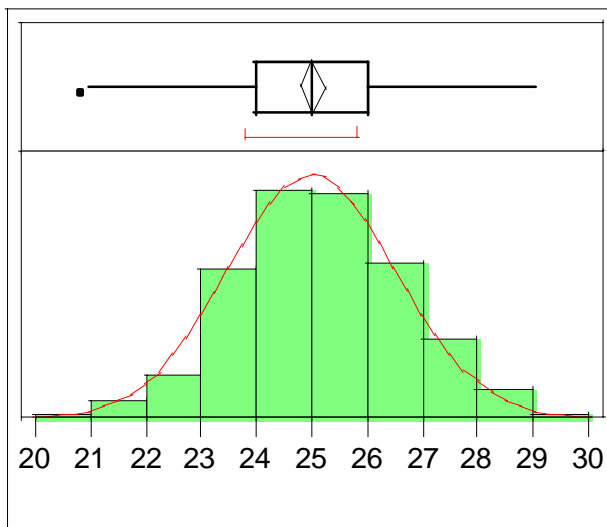


Figure 9: Distribution of means of samples from a symmetric binomial population ( $n = 5$ )

**References**

- [1] Newton, H.J. and Harvill, J.L. (1997). StatConcepts: A visual tour of statistical ideas, Duxbury, CA, pp. 102.
- [2] Serfling, R. J. (2001). Approximation Theorems of Mathematical Statistics, Wiley, pp. 28.
- [3] Yates, D.S., Moore, D.S., and Starnes, D.S. (2003). The Practice of Statistics, Freeman, pp. 521.

University of South Alabama  
Dept of Math & Stat, 307 Univ. Blvd, ILB 325,  
Mobile, AL 36688-0002  
mmulekar@jaguar1.usouthal.edu